*W594–W599* *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* *Published online 25 April 2019*
*doi: 10.1093/nar/gkz289*

# LitSense: making sense of biomedical literature at sentence level

**Alexis Allot†, Qingyu Chen †, Sun Kim†, Roberto Vera Alvarez, Donald C. Comeau, W. John Wilbur and Zhiyong Lu ***

National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received February 08, 2019; Revised April 05, 2019; Editorial Decision April 08, 2019; Accepted April 10, 2019

## ABSTRACT

**Literature search is a routine practice for scientific studies as new discoveries build on knowledge from the past. Current tools (e.g. PubMed, PubMed Central), however, generally require significant effort in query formulation and optimization (especially in searching the full-length articles) and do not allow direct retrieval of specific statements, which is key for tasks such as comparing/validating new findings with previous knowledge and performing evidence attribution in biocuration. Thus, we introduce LitSense, which is the first web-based system that specializes in sentence retrieval for biomedical literature. LitSense provides unified access to PubMed and PMC content with over a half-billion sentences in total. Given a query, LitSense returns best-matching sentences using both a traditional term-weighting approach that up-weights sentences that contain more of the rare terms in the user query as well as a novel neural embedding approach that enables the retrieval of semantically relevant results without explicit keyword match. LitSense provides a user-friendly interface that assists its users to quickly browse the returned sentences in context and/or further filter search results by section or publication date. LitSense also employs PubTator to highlight biomedical entities (e.g. gene/proteins) in the sentences for better result visualization. LitSense is freely available at https://www.ncbi.nlm.nih.gov/research/litsense.**

## INTRODUCTION

Literature search is a routine practice for scientific studies, as new discoveries build on prior knowledge. Indeed, each day millions of users search PubMed (https://pubmed.gov) and PubMed Central (PMC; https://www.ncbi.nlm.nih.gov/pmc) (1), among many others (2–4), seeking answers to their information needs in biomedicine. It is currently impossible, however, for a user to simultaneously query all of the content in both databases with a single search in PubMed or PMC, despite their highly related and even somewhat overlapping content. An earlier study shows that retrieval performance can decrease by more than 10% when PubMed and PMC articles are simply combined for keyword search (5) due to the various differences and complexities in full-text retrieval (6,7). For example, a full-length article often consists of multiple (sub-)topics such that topic shifts are common, which has been shown to result in redundant or irrelevant signals to search engines (7,8).

One solution to address the challenging issues in full-text search is to perform passage- or sentence-level retrieval, as suggested in previous studies (9–12), instead of traditional document-level search due to the unique advantages of the former. For example, sentences have higher locality or information density such that they are more likely to be relevant if they contain multiple query terms. Then, document length is not an issue and retrieval can be more effective. Sentence-level search can play a vital role in a range of biomedical applications, e.g. to quickly compare and contrast new findings with previous knowledge (13), to perform evidence attribution from the literature or to assist biomedical question answering and document summarization (8).

In this paper, we present LitSense, a search system for over a half-billion sentences from the entire 29+ million article abstracts in PubMed and ∼3 million full-text articles in the PMC Text Mining Subset (14). To the best of our knowledge, this is the first sentence search tool that provides unified access to the combined PubMed and PMC contents. In addition, LitSense has several unique features compared to PubMed and PMC. First, in regard to results ranking, LitSense makes use of a state-of-the-art neural embedding approach (15) to traditional term-matching information retrieval (IR) methods for improved performance. Second, LitSense employs PubTator (16), a state-of-the-art bio-entity recognition (NER) tool to highlight bio-entities

*To whom correspondence should be addressed. Tel: +1 301 594 7089; Fax: +1 301 480 2290; Email: zhiyong.lu@nih.gov
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

in search results for better visualization. Third, users can easily locate the position of returned sentences of interest in full text and examine the surrounding text. Other user-friendly features include an intuitive and interactive interface that allows quick browsing of returned sentences, plus sentence filtering by section titles (e.g. Results and Discussion) and publication dates (e.g. last 3 years).

## SYSTEM DESCRIPTION

As shown in Figure 1, LitSense has two major components: 'sentence indexing' and 'sentence search'. We pre-process all sentences from PubMed and PMC and then index them in Solr (http://lucene.apache.org/solr). We also train sent2vec (15), a cutting-edge neural embedding approach, to obtain a semantic representation/vector of each sentence. During search time, for a user query, LitSense first returns sentences that best match the query terms from the Solr database. The retrieved sentences are then re-ranked using semantic vectors. This re-ranked result is displayed to the user in the last step. The following subsections provide a description of our 'sentence indexing' and 'sentence search' components in detail.

### Sentence indexing

We apply multiple text processing steps to extract sentences from PubMed and PMC documents. The first step is 'document filtering' (Figure 1a) to remove lengthy documents such as conference/workshop proceedings in PMC, while keeping only individual research articles. The next step is 'section normalization' (Figure 1b). A full-text article typically comprises multiple sections, such as Introduction, Materials and Methods, Results and so forth. We select only content-relevant sections for further processing and remove sections such as References, Abbreviations, Acknowledgments and Conflict of Interest. After section removal, we assign semantic categories (e.g. Introduction and Results) to the remaining sections by utilizing the BioC repository (14). Note that such semantic categories are also used for filtering retrieved sentences by section titles (Figure 2b). Finally, we split text into sentences, using the PunktSentenceTokenizer from the NLTK toolkit (17) (Figure 1c), which achieved an *F*-score of 98.9% on English newspaper corpora (18). We also remove sentences that are deemed too short or too long (the current thresholds are fewer than 20 characters or longer than 1000 characters), resulting in 611 485 082 sentences as of 4 February 2019.

Next, we index all sentences by Solr for retrieval, and the same data are used for learning semantic representation of sentences, i.e. sentence vectors/embeddings, using sent2vec. Further, we apply PubTator (16) to identify bio-entities (e.g. genes, chemicals and diseases) in these sentences. The parameters used in sent2vec training and the NER performance of PubTator are listed in the Supplementary Data.

### Sentence search

A core function of LitSense is to match a query sentence against a half billion sentences in the corpus. To optimize both retrieval effectiveness and efficiency, we use a two-phase ranking process, similar to PubMed's Best Match algorithm (19). The first phase focuses on efficient retrieval of relevant PubMed and PMC sentences. Specifically, we use the inverse document frequency (IDF) ranking function (20) to find best-matching sentences to the query (Figure 1B), as we observed that IDF performs better than do some other classic models, such as BM25, in our experiments (Table 1). That is, within-sentence term frequency (count of each query word) was not found to be helpful for sentence-level retrieval.

**Table 1.** Performance comparison of BM25, IDF, sent2vec and IDF + sent2vec approaches

| Method | NDCG@1 | NDCG@3 | NDCG@5 |
|---|---|---|---|
| BM25 | 0.5539 | 0.5704 | 0.6070 |
| IDF | 0.6431 | 0.6814 | 0.6944 |
| sent2vec | 0.5479 | 0.5939 | 0.6331 |
| IDF + sent2vec | **0.6919** | **0.6971** | **0.7204** |

We manually labeled about 18 sentences per query for 100 queries and evaluated BM25, IDF, sent2vec and our IDF + sent2vec approaches, using the gold-standard set. NDCG measures the effectiveness of the ranking function by comparing its result with the ideal ranking based on relevance. NDCG@$k$ means the NDCG score at the $k$th rank.

Although term-based matching (i.e. IDF) is efficient for searching a half-billion sentences, it may not fully capture the underlying semantics of natural language (21). That is, two sentences that share the same terms may not be similar in semantics. Recent sentence-embedding approaches aim to address this problem by producing sentence vectors that capture the semantics beyond word level (22,23). Inspired by deep learning, neural sentence-embedding methods have achieved state-of-the-art performance in various sentence-related tasks, i.e. sentence pair semantic similarity (24,25), sentence inference (26) and question entailment (27), and have been shown to be effective in combination with traditional information retrieval methods (28–30). Thus, for the $N$ top-ranked sentences scored by IDF (currently, $N = 100$ for system efficiency), we compute the cosine similarity between the query sentence vector and each retrieved sentence vector, obtained from the previous indexing step. Then, we re-rank the $N$ sentences by averaging IDF and sent2vec similar scores (Figure 1C). Note that we take an average of IDF and sent2vec empirically, following the experiments described below.

We evaluated the performance of our method on a manually annotated dataset that contains 100 query sentences and ∼18 retrieved sentences for each query with relevance labels. Two medical doctors separately annotated each query-sentence pair on a scale of 1 (not relevant) to 5 (highly relevant), with a moderate-to-strong positive Pearson's correlation coefficient of 0.56. Relevance scores from the two annotators were averaged and then used as the final gold standard accordingly. Our dataset and the annotation details are freely available at https://www.ncbi.nlm.nih.gov/research/litsense.

Table 1 shows the performance comparison of BM25, IDF, sent2vec and our IDF + sent2vec approaches using normalized discounted cumulative gain (NDCG) (31). In

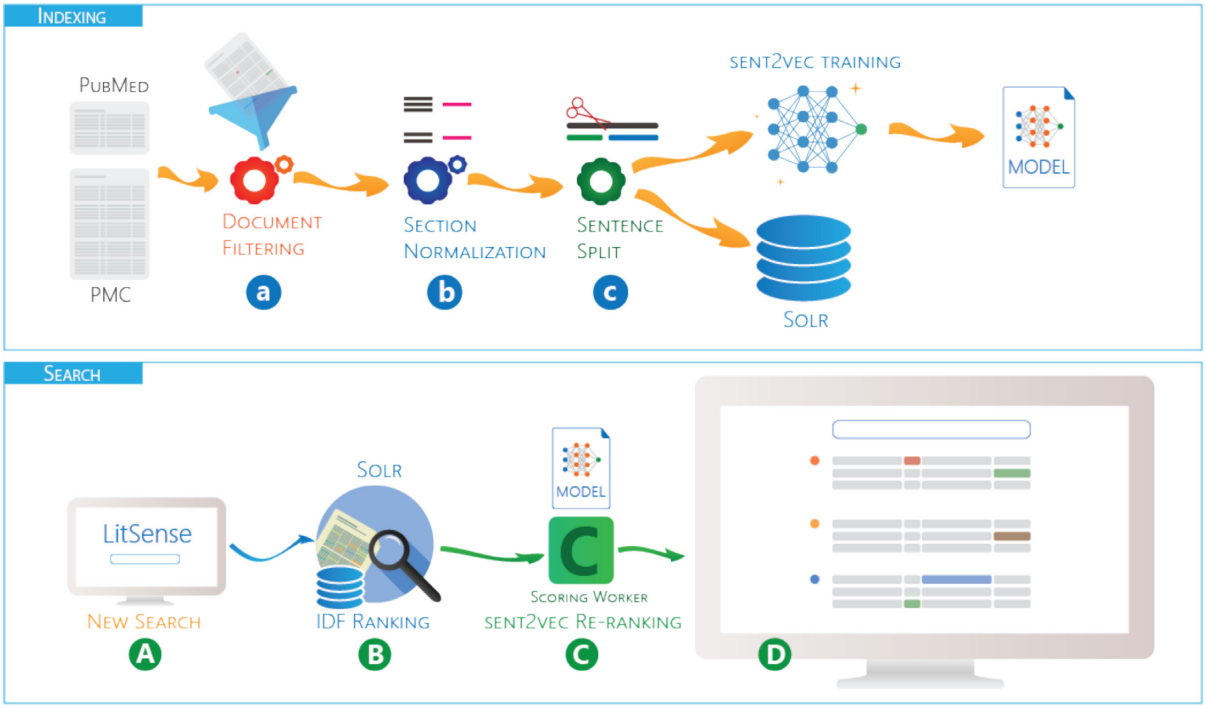


**Figure 1.** System overview. LitSense has two main parts: 'sentence indexing' and 'search'. We first obtain PubMed and PMC documents from the BioC repository (https://www.ncbi.nlm.nih.gov/research/bionlp/APIs). After removing irrelevant documents (a), sections in each article are normalized to semantic categories (b), and then the text is split into sentences (c). The extracted sentences are stored in Solr and used for learning semantic vectors via sent2vec. Given a user query (A), Solr retrieves sentences, using the inverse document frequency (IDF) ranking in Solr (B), and the top-ranked sentences are subsequently re-ranked by semantic vector similarity scores (C). Finally, the system displays the results through the web interface (D).

the table, IDF outperforms BM25, which means that the within-sentence term frequency used in BM25 is not useful for evaluating sentence similarity. Indeed, existing studies note that sentences contain far fewer words than do documents and that counting shared terms between queries and sentences is less effective (32). We also do not find it surprising that IDF performs better than does sent2vec, as semantic similarity may somewhat be limited as a basis for relevant judgments. The last row in Table 1 shows that an improved performance can be obtained by combining the IDF and sent2vec approaches.

### Implementation

We developed LitSense using the Angular framework (https://angular.io) for the frontend and the Django framework (https://www.djangoproject.com) for the backend. We apply stemming to both queries and indexed terms and use synonyms from MeSH (https://meshb.nlm.nih.gov) to improve recall. We fetch PubMed and PMC articles regularly and update our indexing accordingly. LitSense supports the latest version of popular web browsers, such as Chrome, Safari, Firefox and Edge.

## USAGE

LitSense can be accessed through a user-friendly web interface (Figure 2). After a user enters a query sentence in the search bar (Figure 2a), LitSense normalizes the query and retrieves the best-matching sentences, using IDF and sent2vec scores. The results are sorted and displayed in descending order (middle column in Figure 2). Along with retrieved sentences, the likelihood of being relevant (i.e. IDF + sent2vec similarity scores) is shown for each sentence by the colored dot (Figure 2e), ranging from orange (likely to be relevant) to blue (likely to be irrelevant). We also show the provenance of the sentence (Figure 2f). PubMed and PMC IDs are linked to the corresponding PubMed and PMC web pages (Figure 2g). By a single mouse click, users can use the returned sentence as query for a new search (Figure 2h) or display citation information (Figure 2i). 'SEE IN ABSTRACT' or 'SEE IN FULLTEXT' highlights the retrieved sentence in the abstract or full-text article, respectively (Figure 2j). This helps users to easily navigate its surrounding text so that users can get more information, if desired.

Moreover, LitSense provides two useful filtering options. Users can opt to see sentences from certain sections (Figure 2b) and/or sentences from recent years (Figure 2c). The interface also highlights bio-entities in retrieved sentences if they appear also in the query. As explained earlier, bio-entities such as genes, chemicals, diseases, mutations and cell lines are automatically identified by PubTator. The highlights on bio-entities can be toggled on and off by using the BioConcepts menu (Figure 2d).

## USE CASES

Here, we provide examples of how LitSense may be used under real-world circumstances. (The search results presented

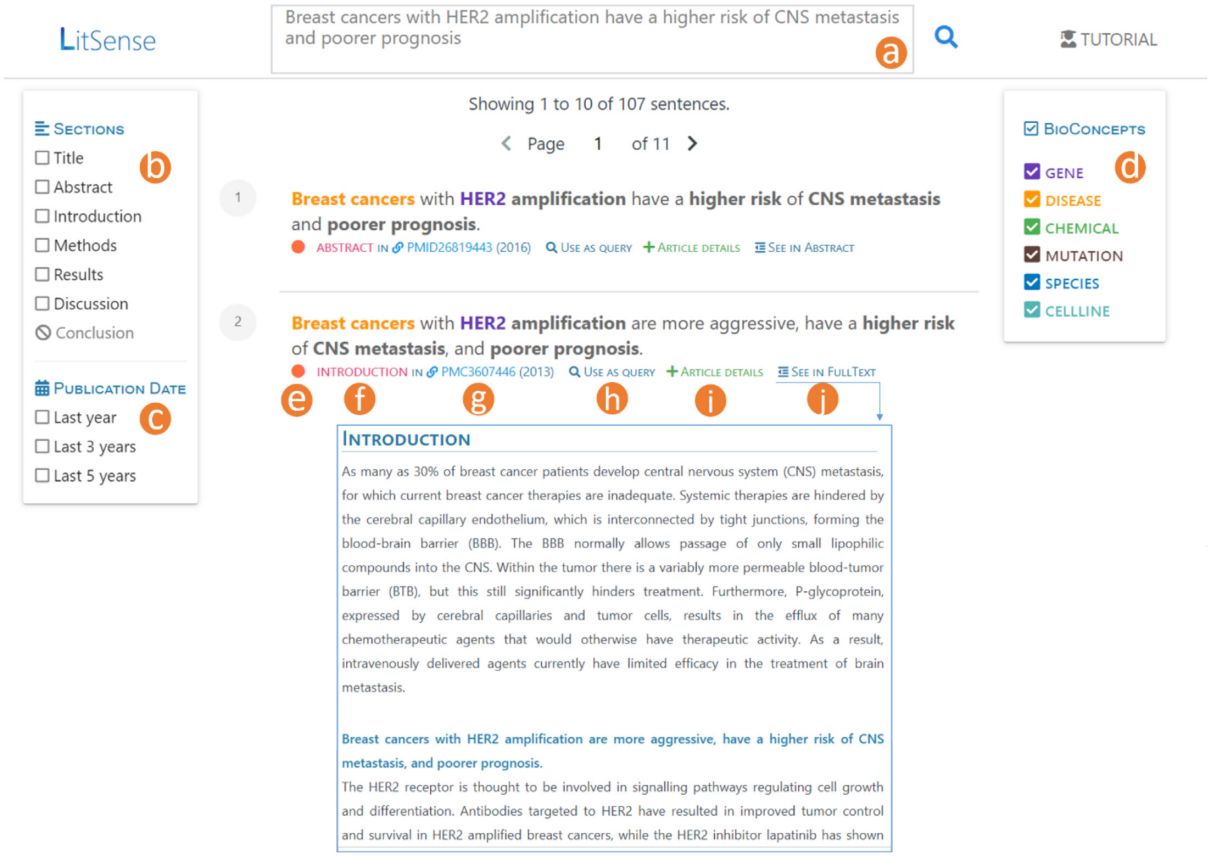


**Figure 2.** LitSense user interface. Users can enter queries into the search bar (a), filter results by a section (b) or a publication date (c), and show/hide the highlights of bio-entities (d). The middle column displays retrieved sentences. The circle icon under each sentence indicates the predicted relevance level from orange (likely to be relevant) to blue (likely to be irrelevant) (e). The same line also shows the provenance of the sentence (f), PubMed/PMC ID (g), a button to use this sentence as query for a new search (h), and citation information (by clicking '+ARTICLE DETAILS') (i). The mouse click on 'SEE IN ABSTRACT/FULLTEXT' opens the entire document with the retrieved sentence highlighted (j).

in this section may be slightly different from those in LitSense online due to our regular system updates.)

### Case 1: Scientists search for similar findings across different studies

Having similar findings from independent studies is common in research and important to assess reproducibility (33). In scholarly publications, key findings are often summarized in a sentence; thus, LitSense can be used to facilitate the search for similar findings.

Here is an example query: 'Autosomal genetic control of human gene expression does not differ across the sexes'. While multiple sentences from the same article (PMCID 5134098) are retrieved by LitSense, other top-ranked sentences include:

- The autosomal genetic control of sexually dimorphic traits in humans is largely the same across the sexes (PMCID 4975899).
- No differences between men and women were found in autosomal genetic control of gene expression (PMCID 5297171).
- Furthermore, our findings suggest that at least one X-linked gene that influences ZPBP2 DMR methylation levels resides on the long arm of chromosome X. This is the first study that attempts dissecting the genetic mechanisms underlying sex-specific differences in methylation levels in a human autosomal region, and our findings may be applicable to other loci (PMCID 5819645).

The specifics in each article might differ, but all of the sentences concern whether there are gender-specific differences in the autosomal genetic control of gene expression.

### Case 2: Biocurators perform evidence attribution

Evidence attribution is an essential step in biocuration workflows (34). For instance, curators in the Conserved Domains Database (CDD) (https://www.ncbi.nlm.nih.gov/Structure/cdd) and UniProt (https://www.uniprot.org) should link a (manually annotated) protein function to the source of evidence (e.g. publications) so that users can trace back to the source and validate the information (35).

The summary of the bZIP Superfamily in CDD (https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=cl21462) starts with the sentence 'Basic leucine zipper (bZIP) factors comprise one of the most important classes of enhancer-type transcription factors'. We use this annotation as a query, assuming that a biocurator wants to link

it to a publication. LitSense ranks the following as the top three similar sentences.

- Dimeric basic leucine zipper (bZIP) factors constitute one of the most important classes of enhancer-type transcription factors (PMID 16731568).
- Basic leucine zipper (bZIP) transcription factors comprise one of the largest gene families in plants (PMID 28955639).
- C/EBPs are members of the basic leucine zipper (bZIP) class of transcription factors (PMCID 2843749).

In CDD, PMID 16731568 is already mapped as a reference article, but the second and third sentences are not linked to the database record. Because the biological databases, such as CDD and UniProt, rely mainly on manual curation, the results obtained from LitSense may facilitate information access, particularly, for the articles that are not referenced in current records.

### Case 3: Non-experts find information from scholarly publications

For people who do not work in medical science, broadcast and digital media are the most common means to get biomedical and clinical news. These non-experts may use PubMed and PMC to seek more academic information, but it is cumbersome to pick out relevant text from full-text articles. Although LitSense was designed for sentence queries, the locality of sentences, explained earlier, may help to identify such information from a set of query keywords.

Assume that one read a news article about a measles outbreak in a certain area; the user might want to try to find information about measles outbreaks and vaccinations. With the query, 'measles outbreak vaccination', LitSense retrieves sentences such as

- Such seasonal outbreak patterns were eliminated in the US after 1981, through the implementation of the highly effective MMR (measles, mumps and rubella) blanket vaccination program (PMCID 5032840).
- Measles outbreaks are highly responsive to vaccination campaigns (PMCID 4885724).
- The measles outbreak in Taiwan seems to have mainly resulted from secondary vaccine failure rather than suboptimal vaccination coverage (PMCID 6144468).

The sentences listed are all relevant to the query and provide a description of different aspects of measles outbreaks and vaccination.

## CONCLUSION

In summary, LitSense provides fast sentence-level retrieval for biomedical literature by including the entire PubMed plus ∼3 million full-text articles in PMC. Further, it allows users to filter retrieved sentences by section and publication date.

LitSense has several known limitations. LitSense is subject to the accuracy of the current text mining tools used for splitting sentences and highlighting entities in the search results, which are known to be imperfect. In addition, LitSense considers that a sentence is a self-sufficient fact, which is not always the case. Some sentences contain distinct factual entities (often separated by ';'), while others make sense only in the context of surrounding sentences. Thus, to improve the quality of our results, we plan to further investigate splitting publications into meaningful and self-sufficient factual entities.

In the future, we also would like to improve the coverage of search results returned by Solr through additional synonymy. Finally, because sentences frequently use pronouns or other terms to reference entities that are outside the sentence, we plan to address such issues using anaphora resolution.

## DATA AVAILABILITY

LitSense is free and open to all users and there is no login requirement. LitSense can be accessed at https://www.ncbi.nlm.nih.gov/research/litsense.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors would like to thank David Landsman, Won Kim, Tiarnan Keenan and Mingzhang Yang for their help with implementing the LitSense interface. They also would like to thank Benjamin S. Wilbur, MD, for annotating the evaluation dataset.

## FUNDING

Intramural Research Program of the National Library of Medicine, National Institutes of Health. Funding for open access charge: Intramural Research Program of the National Library of Medicine, National Institutes of Health.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Fiorini,N., Leaman,R., Lipman,D.J. and Lu,Z. (2018) How user intelligence is improving PubMed. *Nat. Biotechnol.*, **36**, 937–945.
2. Jensen,L.J., Saric,J. and Bork,P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.
3. Europe PMC Consortium. (2015) Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Res.*, **43**, D1042–D1048.
4. Doms,A. and Schroeder,M. (2005) GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Res.*, **33**, W783–W786.
5. Kim,W.G., Yeganova,L., Wilbur,W.J. and Lu,Z. (2018) MeSH-based dataset for measuring the relevance of text retrieval. *Proceedings of the BioNLP 2018 Workshop*. Melbourne, Australia, pp. 161–165.
6. Cohen,K.B., Johnson,H.L., Verspoor,K., Roeder,C. and Hunter,L.E. (2010) The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, **11**, 492.
7. Lin,J. (2009) Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, **10**, 46.
8. Sarrouti,M. and El Alaoui,S.O. (2017) A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering. *J. Biomed. Inform.*, **68**, 96–103.

9. Kaszkiel,M. and Zobel,J. (1997) Passage retrieval revisited. *ACM SIGIR Forum*, **31**, 178–185.
10. Blanco,R. and Zaragoza,H. (2010) Finding support sentences for entities. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Geneva, Switzerland, pp. 339–346.
11. Losada,D.E. and Fernández,R.T. (2007) Highly frequent terms and sentence retrieval. *Proceedings of the International Symposium on String Processing and Information Retrieval*. Santiago de Chile, Chile, pp. 217–228.
12. Hersh,W. and Voorhees,E. (2009) TREC genomics special issue overview. *Inform Retrieval*, **12**, 1–15.
13. Wallach,J.D., Boyack,K.W. and Ioannidis,J.P. (2018) Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *PLoS Biol.*, **16**, e2006930.
14. Comeau,D.C., Wei,C.-H., Doğan,R.I. and Lu,Z. (2019) PMC text mining subset in BioC: about 3 million full text articles and growing. *Bioinformatics*, doi:10.1093/bioinformatics/btz070.
15. Pagliardini,M., Gupta,P. and Jaggi,M. (2018) Unsupervised learning of sentence embeddings using compositional n-gram features. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana, Vol. **1**, pp. 528–540.
16. Wei,C.-H., Kao,H.-Y. and Lu,Z. (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.*, **41**, W518–W522.
17. Loper,E. and Bird,S. (2002) NLTK: the Natural Language Toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Philadelphia, Pennsylvania, Vol. **1**, pp. 63–70.
18. Kiss,T. and Strunk,J. (2006) Unsupervised multilingual sentence boundary detection. *Comput. Linguist.*, **32**, 485–525.
19. Fiorini,N., Canese,K., Starchenko,G., Kireev,E., Kim,W., Miller,V., Osipov,M., Kholodov,M., Ismagilov,R., Mohan,S. *et al.* (2018) Best Match: new relevance search for PubMed. *PLoS Biol.*, **16**, e2005343.
20. Sparck Jones,K. (1972) A statistical interpretation of term specificity and its application in retrieval. *J. Doc.*, **28**, 11–21.
21. Onal,K.D., Zhang,Y., Altingovde,I.S., Rahman,M.M., Karagoz,P., Braylan,A., Dang,B., Chang,H.-L., Kim,H. and McNamara,Q. (2018) Neural information retrieval: At the end of the early years. *Inform. Retrieval J.*, **21**, 111–182.
22. Ramaprabha,J., Das,S. and Mukerjee,P. (2018) Survey on sentence similarity evaluation using deep learning. *J. Phys. Conf. Ser.*, **1000**, 012070.
23. Xie,Y., Le,L., Zhou,Y. and Raghavan,V.V. (2018) Deep learning for natural language processing. *Handbook of Statistics*. Elsevier, Amsterdam, Vol. **38**, pp. 317–328.
24. Cer,D., Diab,M., Agirre,E., Lopez-Gazpio,I. and Specia,L. (2017) SemEval-2017 Task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv doi: https://arxiv.org/abs/1708.00055, 31 July 2017, preprint: not peer reviewed.
25. Chen,Q., Peng,Y. and Lu,Z. (2018) BioSentVec: creating sentence embeddings for biomedical texts. arXiv doi: https://arxiv.org/abs/1810.09302, 22 October 2018, preprint: not peer reviewed.
26. Poliak,A., Haldar,A., Rudinger,R., Hu,J.E., Pavlick,E., White,A.S. and Van Durme,B. (2018) Collecting diverse natural language inference problems for sentence representation evaluation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 67–81.
27. Dasgupta,I., Guo,D., Stuhlmüller,A., Gershman,S.J. and Goodman,N.D. (2018) Evaluating compositionality in sentence embeddings. arXiv doi: https://arxiv.org/abs/1802.04302, 12 February 2018, preprint: not peer reviewed.
28. Hoogeveen,D., Wang,L., Baldwin,T. and Verspoor,K.M. (2018) Web forum retrieval and text analytics: a survey. *Found. Trends Inform. Retrieval*, **12**, 1–163.
29. Gupta,V., Chinnakotla,M. and Shrivastava,M. (2018) Retrieve and re-rank: A simple and effective IR approach to simple question answering over knowledge graphs. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium, pp. 22–27.
30. Das,A., Yenala,H., Chinnakotla,M. and Shrivastava,M. (2016) Together we stand: Siamese networks for similar question retrieval. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, Vol. **1**, pp. 378–387.
31. Järvelin,K. and Kekäläinen,J. (2002) Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inform. Syst. (TOIS)*, **20**, 422–446.
32. Murdock,V.G. (2006) *Aspects of Sentence Retrieval*. University of Massachusetts Amherst.
33. Goodman,S.N., Fanelli,D. and Ioannidis,J.P. (2016) What does research reproducibility mean? *Sci. Transl. Med.*, **8**, 341ps312–341ps312.
34. Lu,Z. and Hirschman,L. (2012) Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database*, **2012**, bas043.
35. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M., Bansal,P., Bridge,A.J., Poux,S., Bougueleret,L. and Xenarios,I. (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Plant Bioinformatics*, **1374**, 23–54.